REMARKS

The claims are 1-20. Claims 1, 3, 4, 7, 14, 16 and 18 have been amended. Claims 1, 3, 7, 14, 16 and 18 are in independent form. Favorable reconsideration and allowance of the subject application are respectfully requested in view of the following comments.

Claims 1, 3, 7, 14, 16 and 18 have been amended to clarify that the energy bar claimed has about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size. Support for the amendment can be found, for example, in paragraph [0016] on pages 4-5, and paragraph [0042] on page 12 of the specification.

Claim 4 has been amended to correct a minor error.

Claims 1-3, 7, 15, 17, 19 and 20 stand rejected under 35 U.S.C. § 112, second paragraph, for allegedly being indefinite. Specifically, the Office Action has objected to the use of the terms "hedonic score," "confidence level," and "acceptability." Applicants respectfully direct the Examiner's attention to paragraphs [0020] and [0022] on pages 5 and 6 of the specification, where the definitions for "hedonic score" and "confidence level" are provided. Moreover, Applicants note that the term "acceptability" is understood in the food industry to denote a consumer's willingness to eat a product. *See* Principles of Sensory Evaluation of Food, 1965, p. 278. Applicants also wish to point out that one skilled in the art understands that the hedonic score and confidence intervals are statistically determined measurements and are reproducible within a certain degree of error. Applicants respectfully direct the Examiner's attention to the following publications, which demonstrate the use of

these terms throughout the food industry: Sensory Analysis of Foods, pp. 250, 254-257, and 366[1]; Statistical Methods in Food and Consumer Research, pp. 7 and 8; and Principles of Sensory Evaluation of Food pp. 275-289. Copies of each are enclosed for the Examiner's convenience. Accordingly, Applicants respectfully request withdrawal of the Section 112 rejections.

Claims 1-20 have been provisionally rejected under the judicially created doctrine of obviousness-type double patenting as being unpatentable over claim 20 of copending Application No. 10/272,571 (the '571 application) and claims 1-20 of copending Application No. 10/271,710 (the '710 application). Applicants note that the '571 application was abandoned on September 15, 2004 for being non-responsive to the Office Action issued on June 15, 2004. As such, the provisional rejection based on the '571 application is rendered moot. Regarding the provisional rejection based upon the '710 application, a Terminal Disclaimer is submitted herewith. In light of the above comments, it is believed that the provisional double patenting rejections have been obviated, and their withdrawal is therefore respectfully requested.

Claims 1-10 stand rejected under 35 U.S.C. § 102(b) as allegedly being anticipated by U.S. Patent No. 4,055,669 ("*Kelly*"). Claims 1-13 and 18-20 stand rejected under 35 U.S.C. § 103(a) as allegedly being obvious over Kelly in view of U.S. Patent No. 6,592,915 ("*Froseth*") and a recipe for Pfeffernusse found in the book titled, Joy of Cooking ("*Rombauer*"), on page 708. Claims 14-17 stand rejected under 35 U.S.C. § 102(b) as allegedly being anticipated by *Rombauer*. Applicants respectfully traverse these rejections, in view of the comments set forth below.

---

[1] The hedonic score may be based on a nine-point scale or seven-point scale. For purposes of the present

As amended, claim 1 is directed to an energy food bar that provides about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size.

Kelly is directed to a high protein fat occluded food composition made of cereal particles and a binder. The binder includes a protein source coated with an edible fat, which masks the protein flavor, making the binder taste bland.

Applicants have reviewed *Kelly* and have determined that the amount of fat in the food composition exceeds the permissible amount set forth in claim 1 of about 2 to about 10 g of fat, based on a 55 g serving size. In column 2, lines 56-58, *Kelly* discloses that a binder composition makes up 60-70% of the food composition. *Kelly* further states in column 3, lines 61-64, that "[t]he fat content of the binder composition ranges from a minimum of about 33% by weight to a maximum of about 85% by weight, preferably about 47% by weight[.]" Therefore, the minimum amount of fat present in the binder composition of *Kelly* can be calculated by multiplying the (percent binder) by the (percent fat in the binder) by the (serving size). For a 55 g serving, the minimum amount of fat present in the binder composition alone is 10.9 g of fat (55 g X (33% fat) X (60% binder)). Moreover, additional fat in the food composition of *Kelly* is found in the cereal components that make up the other 40% of the food composition. Low fat cereal components such as crisp rice or corn flakes have about 0.5% fat. For a 55 g serving basis, this would amount to 0.1 g of fat (55 g X (0.5% fat) X (40% cereal)) in the cereal portion. The minimum total amount of fat in the food composition is therefore calculated to be 11 g of fat. This clearly exceeds the range of about 2

---

invention, a seven-point scale was selected.

to about 10 grams of fat permitted in the energy food product set forth in claim 1. As such, it is respectfully submitted that claim 1 is patentable over *Kelly*.

Claim 2 directly depends from claim 1. For at least the same reasons discussed above in connection with claim 1, claim 2 is patentable over *Kelly*.

Independent claims 3 and 7, as well their respective dependent claims, require that the energy bar have about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, and about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size. As such, claims 3 and 7 and their respective dependent claims, are patentable over *Kelly*.

Froseth discloses a layered cereal bar having identifiable ready to eat cereal pieces and at least one visible filling layer. The cereal bar has a total nutrient level equal to or greater than the nutrient level of a single serving of boxed cereal with milk.

Froseth, however, does not disclose a cereal bar having about 1 to about 5 g of fortification components. In column 15, lines 17-25, *Froseth*, discloses an embodiment where the amount of tricalcium phosphate (TCP), i.e., mineral, in the binder is 3% on a weight basis. Froseth also discloses that the binder makes up 40% of the cereal bar (*see* column 11, lines 15-16). For a 55 g serving basis, the amount of TCP in the cereal bar can be calculated to be 0.66 g of TCP (55 g X (40% binder) X (3% TCP in binder)). Therefore, the cereal bar of *Froseth* does not fall within the fortification component range of about 1 to about 4.5 grams in the energy bar set forth in claim 1. As such, the cereal bar of *Froseth* would not qualify as an energy bar.

Rombauer is cited for disclosing a recipe for Pfeffernusse. The Office Action states that "an energy matrix made of corn syrup which is combined with a solid component, grated lemon rind, which is mixed into a fat-carbohydrate matrix (butter and sugar)(page 708). The composition is considered to have a lubricious mouthfeel since the claimed ingredients are used."

Applicants note, however, that *Rombauer* fails to meet the protein level required by the range of about 5 to about 40 g, set forth in claim 1. The table below provides a breakdown of the ingredients used to make the Pfeffernusse composition.

PFEFFERNUSSE

Ingredient		Grams of Protein (based on 55 g serving)
Flour	2.01 cups	3.21
Baking Powder	0.75 tsp	
Baking Soda	0.13 tsp	
Salt	0.25 tsp	
Black Pepper	0.25 tsp	0.01
Nutmeg	0.25 tsp	0.01
Cinnamon	1 tsp	0.01
Fennel Seed	1 tsp	0.05
Butter	0.5 cups	0.03
Sugar	0.33 cup	
Egg	1	0.47
Chopped Almonds	0.25 cup	0.82
Chopped Citron	1 tbsp	
Orange Peel	0.25 cup	
Molasses	0.33 cup	
Corn Syrup	1 tbsp	
Brandy	0.33 cup	
Lemon Rind	1 tsp	
Lemon Juice	1 tbsp	
TOTAL		4.61

Applicants have determined that the protein content in the Pfeffernusse composition is approximately 4.6 g. This does not fall within the protein range of about 5 to

about 40 g (based on a 55 serving), claimed in claim 1. Moreover, the Pfeffernusse composition is not seen to include fortification components. As such the range of about 1 to about 4.5 g of fortification components, set forth in claim 1 is not met. Clearly, the Pfeffernusse composition of *Rombauer*, does not qualify as an energy bar.

Applicants respectfully submit that *Kelly*, *Froseth*, and *Rombauer*, whether taken alone or in any permissible combination, do not disclose or suggest the presently claimed invention of an energy bar that provides about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size, as set forth in claim 1.

Claim 2 directly depends from claim 1. For at least the same reasons discussed above in connection with claim 1, claim 2 is patentable over *Kelly*, *Froseth*, and *Rombauer* whether considered alone or in any permissible combination.

Like claim 1, independent claims 3, 7 and 18 each require that the energy bar have about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, and about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size. For at least the same reasons discussed above for claim 1, claims 3, 7 and 18 are patentable over *Kelly*, *Froseth*, and *Rombauer*, whether considered alone or in combination.

Claim 14 is a product by process claim and claim 18 is a method claim, which require that the energy bar have about 2 to about 55 g of carbohydrates, about 1 to about 4.5 g of fortification components, about 5 to about 40 g of protein, about 2 to about 10 g of fat, and

about 150 to about 300 calories, and a moisture content of less than about 15% by weight, based on a 55 g serving size.

As previously noted, the *Rombauer* Pfeffernusse composition has approximately 4.6 g of protein (based on a 55 g serving) and no fortification components. Therefore the Pfeffernusse composition does not meet the protein level of about 5 g to about 40 g of protein, and the fortification level of about 1 to about 4.5 g, set forth in claims 14 and 16. As such, claims 14 and 16 are patentable over *Rombauer*.

Claim 15 depends from claim 14, and claim 17 depends from claim 16. Claims 15 and 17 are also patentable over *Rombauer* for the same reasons discussed above for claims 14 and 16.

In view of the foregoing remarks, Applicants respectfully request favorable reconsideration and early passage to issue of the present application.

Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our below listed address.

Respectfully submitted,

Attorney for Applicants
Victor Tsu
Registration No. 46,185

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, NY 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 492806v1

In re Application of:

EDWARD L. RAPP ET AL.

Application No.: 10/615,249

Filed: July 8, 2003

For: TASTING ENERGY BAR
(As Amended)

Docket No. 02280.003720.

Examiner: H. F. Pratt

Group Art Unit: 1761

Date: April 4, 2005

THE COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Transmitted herewith is an Amendment and a Terminal Disclaimer in the above-identified application.

[X] No additional fee is required.

The fee has been calculated as shown below

CLAIMS AS AMENDED						
	(2) CLAIMS REMAINING AFTER AMENDMENT		(4) HIGHEST NO. PREVIOUSLY PAID FOR	(5) PRESENT EXTRA	RATE	ADDITIONAL FEE
TOTAL CLAIMS	* 20	MINUS	** 20	= 0	x $25 $50	0.00
INDEP. CLAIMS	* 6	MINUS	*** 6	= 0	x $100 $200	0.00
Fee for Multiple Dependent claims $180°/$360						
			TOTAL ADDITIONAL FEE FOR THIS AMENDMENT---			0.00

* If the entry in Column 2 is less than the entry in Column 4, write "0" in Column 5.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, write "20" in this space.
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, write "3" in this space.

[ ] Verified Statement claiming small entity status is enclosed, if not filed previously.

[ ] A check in the amount of $______________ is enclosed.

[ ] Charge $______________ to Deposit Account No. 06-1205. A duplicate copy of this sheet is enclosed.

[X] Any prior general authorization to charge an issue fee under 37 C.F.R. 1.18 to Deposit Account No. 06-1205 is hereby revoked. The Commissioner is hereby authorized to charge any additional fees under 37 C.F.R. 1.16 and 1.17 which may be required during the entire pendency of this application, or to credit any overpayment, to Deposit Account No. 06-1205. A duplicate copy of this paper is enclosed.

[ ] A check in the amount of $___________ to cover the fee for a _________ month extension is enclosed.

[X] A check in the amount of $ 130.00 to cover the Terminal Disclaimer fee is enclosed.

[X] Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our address given below.

Respectfully submitted,

Victor Tsu

Victor Tsu
Attorney for Applicants
Registration No.: 46,185

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3800
Facsimile: (212) 218-2200

Form #120

NY_MAIN 492639v1

CHNOLOGY
hs

C. O. CHICHESTER
University of California
Davis, California

G. F. STEWART
University of California
Davis, California

---

ward B. Roessler, PRINCIPLES

ING (*in preparation*)

PRINCIPLES OF SENSORY EVALUATION OF FOOD

by

Maynard A. Amerine
Rose Marie Pangborn
Edward B. Roessler

DEPARTMENTS OF VITICULTURE AND ENOLOGY,
FOOD SCIENCE AND TECHNOLOGY, AND MATHEMATICS,
UNIVERSITY OF CALIFORNIA,
DAVIS, CALIFORNIA

1965

ACADEMIC PRESS New York and London

ACADEMIC PRESS INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 65-22766

PRINTED IN THE UNITED STATES OF AMERICA.

Food science deals w
viding food for human c
harvesting to serving. I
involve biochemistry, mi
basic sciences, as well a
other applied sciences.
been primarily on econ
nutritious foods. Unive
world have concerned t
nutritive composition, n
tional properties of foo
World War II focus
that foods were someti
how sound and nutritio
gradually changed the
new and cheaper meth
quently altered the sen
phasized the growing n
—the sensory analysis
reveals the rapid grow
natural that, in 1957,
an upper-division cours
foods by sensory meth
course.

Our philosophy has
analysis of foods rests
and an understanding
addition is careful stat
new understanding of s
tion with physical and

This text therefore
chology of the senses,
ology, and appropriate
of measuring consumer
clude a brief treatment
istics and various phys
belief that objective te
subjective methods use
food acceptance and p
so it is imperative that

We wish to thank,

Chapter 6

Laboratory Studies: Types and Principles

Foods are submitted to sensory examination to provide information that can lead to product improvement, quality maintenance, the development of new products, or analysis of the market. This section summarizes the most important types of sensory problems encountered by food research groups and the main types of procedures used in solving them. This chapter covers the use of laboratory panels, as do Chapters 7 and 8. Consumer testing is discussed in Chapter 9, and statistical procedures for evaluation of the results of both types of panels are covered in Chapter 10.

Tests may be conducted to: (1) select qualified judges and study human perception of food attributes; (2) correlate sensory with chemical and physical measurements; (3) study processing effects, maintain quality, evaluate raw material selection, establish storage stability, or reduce costs; (4) evaluate quality; or (5) determine consumer reaction. Each of these purposes requires appropriate tests. In general, laboratory panels are used for the first three purposes, highly trained experts for the fourth, and large consumer groups for the last.

In this text we distinguish between two types of laboratory panels: (1) those which determine simple differences between treated samples; and (2) those which determine directional differences. Both are laboratory panels, and sometimes untrained judges are used, but it is the thesis of this book that trained subjects are more useful. The advantages of such panels are discussed in Chapter 7.

I. Types of Tests

The most important types of tests and their utilization are briefly described here. More detailed information of each procedure is given in Chapters 7, 8, and 10.

A. Difference Tests

The common true difference tests are referred to as single-stimulus, paired-stimuli, duo-trio, triangle, and multi-sample tests. In tests which

do not reveal statistically significant differences between treatments, no further evaluation is needed. When differences are found, however, directional difference tests are used to establish the nature and magnitude of difference. After a significant difference has been established by a laboratory panel, consumers may be asked to express preferences.

Since most perceptual judgments are relative, *single-sample* presentation is used infrequently, except at the consumer level. Expert tasters of wines, beers, coffee, tea, and dairy products rate single samples, but they evaluate the quality of many samples at a time and compare them against their pre-established "memory standard." Occasionally a method called "A–not A" is used (Peryam, 1958), in which a standard, A, is presented followed by one or more coded samples. The judge indicates which one(s) *is* (are) A. This method may be classified as a paired comparison rather than single presentation since each coded sample is compared with the standard.

In the *paired-stimuli* procedure, judges simply specify whether there is a difference between two samples. When the judge also indicates what sensory characteristic distinguishes the two samples, we speak of the test as a *paired-comparison*. The samples are presented in a counter-balanced design, and a forced-choice is usually required. One half of the responses could be correct due to chance alone. The number of samples tested at a single session will depend on the commodity, the experience of the judges, and the amount of time and sample available. Paired testing is typically used in comparing new with old processing procedures, in quality control, and in preference testing at the consumer level.

The *duo-trio* is a modified paired presentation in which one sample is identified and presented first, followed by two coded samples, one of which is identical with the standard. The judge is asked which of the two is the same as the first sample. This method is primarily a laboratory tool for use with trained subjects. It lends itself to use for quality control and for selection of judges of superior discrimination.

In the *triangle test*, two identical and one different samples are presented simultaneously and the judge is asked to indicate the odd sample. Correct identification due to chance alone is one third. Like the duo-trio method, the triangle test should be used only by trained laboratory judges, and is suited to similar problems.

B. Rank Order

Ranking is used to determine how several samples differ on the basis of a single characteristic. A group of coded samples (which may contain a control, or standard) are presented simultaneously, and the judge is asked to rank them in order of the intensity of a specified characteristic.

treatments, no
however, direc-
d magnitude of
ablished by a
eferences.
mple presenta-
xpert tasters of
mples, but they
compare them
nally a method
standard, A, is
judge indicates
s a paired com-
sample is com-

whether there
indicates what
peak of the test
unter-balanced
of the responses
ples tested at a
erience of the
aired testing is
procedures, in
level.
a one sample is
amples, one of
hich of the two
laboratory tool
ity control and

mples are pre-
ne odd sample.
ke the duo-trio
ned laboratory

er on the basis
h may contain
d the judge is
characteristic.

This method is suitable for use by laboratory judges in product or process evaluation, by experts for selecting the best sample for a particular use, and by consumers for expressing relative acceptability among a limited number of samples. It is of importance that all judges use the same criteria. When necessary, one criterion (sweetness, for example) can be ranked, after which another criterion (sourness, viscosity, etc.) may be ranked in another set of the same samples.

C. Scoring Tests

The best use of *scoring* tests is in comparisons of a control sample with several experimental samples. The scoring may be expressed in terms of deviation from a reference—"no difference from control" to "very large difference from control." In other experiments, scores may be used on an absolute basis *if the scale is clearly defined and understood by all judges.* Although difference-from-control tests have been used widely by laboratory panels, the results may be meaningless if the judges change the basis of their scoring as the test proceeds, i.e., judges become experts. Thus, this method is best suited for use by experts. The test may be administered to consumers if it is clearly explained and the decisions required are simple.

Tests in which deviation from a control is measured are used for product or process evaluation and critical tests on basic perception of sensory attributes. Scoring tests are also used in new-product development, quality control, storage stability tests, screening of intensity levels, and measuring judge characteristics such as leniency, reproducibility, or central tendency (see Chapter 5).

D. Descriptive Tests

Descriptive sensory analyses are best conducted only by highly trained experts completely familiar with the product or the process. Such tests are used effectively in new-product development, in product or process improvement, for quality control, and for training judges for future testing. One type of descriptive test—hedonic—in which degree of liking is described, is suitable at the consumer level. Among the types of descriptive tests currently in use are scalar scoring of various types, hedonic ratings, semantic differential tests, and Arthur D. Little's "Flavor Profile" (see Chapter 8, Section V).

E. Hedonic Scaling

Scoring is called hedonic when the judge expresses his degree of liking by checking a point on a scale ranging from extreme disapproval to extreme approval. A five- to nine-point balanced scale is usually em-

ployed. Hedonic ratings are converted to scores and treated by rank analysis or analysis of variance. As indicated above, this test has been used both by experts and by untrained consumers, but we feel it is more effectively applicable to the latter.

F. Acceptance and Preference

Distinction should be made between acceptance, which is a willingness to use or eat a product, and preference, which relates to a greater degree of acceptance of one product over another when a choice is presented. The acceptance or preferences of a laboratory panel are of very limited value except in gross screening of treatments. Some of the test methods described above can be adapted to measurement of consumer reaction (see Chapter 9).

G. Other Methods

Dilution tests, described in Chapter 9, have been used for laboratory testing of selected treatments, employing methods of presentation described above, i.e., single, paired, and multiple samples. *Threshold* tests are seldom used except in studies where it is desirable to establish the minimum detectable difference of an additive or of an off flavor. Threshold and dilution tests have been used to a limited extent to select judges who can detect specific sensory properties. When so used, the test materials and their concentrations should be the same as those likely to be encountered in the actual test. Sequential analysis (Chapter 10) can be used to analyze the results.

It is our belief that laboratory judges should be carefully selected and screened on the basis of their sensitivity to the differences that may be encountered in the experimental samples. In this sense, all laboratory panels should consist of experts. It is recognized that in many organizations the time, money, and personnel necessary to achieve this goal are unavailable, but unless judges have had extensive training and experience, they should not be expected to make meaningful evaluations of quality, particularly of a descriptive nature. Neither should a laboratory panel, whether small or large, experienced or inexperienced, presume to predict consumer acceptance or preference. Preferences of a laboratory group are representative only of a limited and unknown portion of the consuming public. This concept is discussed in considerable detail in Chapter 9.

II. Panel Selection and Testing Environment

Systematic analysis of the sensory properties of foods involves the use of human subjects in a laboratory environment. The sensitivity and re-

producibility of the analytical tool (in this case, the judge) greatly influence the direction and validity of the results. The environment under which the judgments are obtained also influences the data. Of additional importance are the time and labor and the supplies and equipment involved, for these factors materially control the cost of sensory analyses. We agree with Foster (1954) that more emphasis must be placed on controlling physical and psychological influences in sensory testing of foods. Unfortunately, the data available for a wide variety of food types are not adequate for the determination of the optimum ranges for all variables.

A. Panel Selection

There is considerable controversy in the literature on the value of a sensory panel that has been selected and trained. Much of the confusion has arisen because discrimination or difference tests have not been distinguished from quality or consumer types of studies. In some cases a failure to find differences between trained and untrained panels in ability to discriminate has had its origin in methodological or statistical deficiencies. Tarver and Ellis (1961) believe the following considerations are important in selecting judges for flavor-difference tests: (1) precision or inherent ability to duplicate a difference judgment; (2) reliability or absence of bias in detecting a flavor difference; and (3) a tolerance level or inherent sensitivity to a particular flavor difference. According to Kramer *et al.* (1961), if the simulation of consumer reaction is the sole aim, a trained panel is not needed and should be avoided. In some cases it may be important to select individuals who are superior in their ability to detect differences. It is difficult, if not impossible, with our present lack of knowledge of consumer response, to select panels that will show good agreement with consumer evaluation. The problem seems to be our inability to define the difference and to train the panel to recognize the difference. Furthermore, the consumer uses many criteria other than sensory in evaluating foods.

Various procedures, based on intuition, rational judgment, or experimentation, have been applied in selecting people whose performance in sensory tests will be superior to that of an unselected population (Dawson *et al.*, 1963). These methods have been tested with varying degrees of success. One major problem is the amount of pretesting work required to establish reliable selection. A further difficulty may be an experimenter's inability to specify accurately the nature of the panel member's task. "Quickie" methods of panel selection, based upon only a few tests, have generally not been very satisfactory. On the other hand, although the tedious process of selecting subjects on the basis of sensitivity to the

basic tastes is often recommended, the method is of doubtful value (Mackey and Jones, 1954; Peryam, 1958).

Since randomly selected and untrained individuals are variable in their judgments, large panels are needed for results that are stable and sensitive. By selecting the most stable and sensitive members and training them, one might expect to obtain a small but efficient panel. Selection is important since individuals differ considerably in sensitivity, interest, motivation, and ability to judge differences. Discriminatory skill need not be general; a good wine taster may not be a good judge of chocolates. Girardot *et al.* (1952) found that candidates who did well on some products often did poorly on others. Seldom is a judge equally proficient in testing all qualities and all flavors of foods. The skill of a connoisseur has been attributed to knowledge of what signs to look for and how to interpret them rather than to increased sensitivity to stimuli (Metzner, 1943). An ability or aptitude for flavor assessment could conceivably vary in three ways: between individuals, between products, and at different times for the same individuals and products (see Coppock *et al.*, 1952; Harvey, 1953). Thus it is evident that a general-purpose panel will be less useful than a specific panel selected for the product and method being tested. A general-purpose panel could be used for gross screening, however, when precision must be sacrificed to save time and expense. A sensory panel should be considered as a tool, and, as such, it can be compared to suitable chemical methods (Lowe and Stewart, 1947). Certain methods and tools may be used to show gross differences, but, as the measurements needed become more refined and precise, the methods and tools required for accurate sensory testing become more sensitive.

Moser *et al.* (1950) considered that selection and training of judges on the basis of sensitivities and consistencies are of extreme importance in evaluating edible oils. In selecting panels, those investigators used a double elimination test (see Chapter 6, Section II,C) based on acuity in oil evaluation. In scoring bitterness in orange juice, Coote (1956) illustrated the necessity of careful training and selection of panels for estimating the degree of bitterness. For beer-tasting tests, Helm and Trolle (1946) selected 20 out of 90 prospective judges. These 20 had the highest percentages of correct selections in triangle tests and were considered to compose a far more suitable taste panel than the original group. Kirkpatrick *et al.* (1957) showed the importance of panel selection for evaluation of milk and biscuits.

Any method of selection should include a preliminary training period designed to acquaint the tasters with the quality factors involved in the product to be tested. This should be followed by a blind test designed to

show the individual's
and Elder, 1950).

B. SCREENING

Most investigators
ing panel members, ir
differences between s
tion; (2) ability to re
parison with other pa
ences in samples to b
the extent to which s
ance in actual tests.

Kramer *et al.* (19
ficient for selecting p
tecting flavor differen
who performed best o
average of the origina
more efficient group. I
have resulted in a still

A general approach
as test materials the sa
tests to obtain variatio
met with in the actua
so that the group as a
individuals will fail; (
later; (5) start with a
a selection test that is
quired; (6) screen on
a top-ranking group of
at each stage reject th
people than will be re
tine task; it requires ju
criteria of achievemen
selection. According to
tion assumes a good p
be perfect.

It is felt that a pers
the skill he has develop
he may note and detec
enced judge. He can
and usually has a be
employed.

show the individual's relative perception and discrimination (Harrison and Elder, 1950).

B. SCREENING

Most investigators employ some type of screening process for selecting panel members, including specific tests based on: (1) discriminating differences between solutions or substances of known chemical composition; (2) ability to recognize flavors or odors; (3) performance in comparison with other panel members; and (4) ability to discriminate differences in samples to be used later in the test. The pertinent question is the extent to which selection devices are reflected in superior performance in actual tests.

Kramer *et al.* (1961) reported that a single screening was insufficient for selecting panel members of continued superior ability in detecting flavor differences. After a first screening of 28 candidates, the 12 who performed best originally did not perform more efficiently than the average of the original 28 candidates. A second screening resulted in a more efficient group. Further screening and training would undoubtedly have resulted in a still more efficient panel.

A general approach may be summarized, stepwise, as follows: (1) use as test materials the same product that will be tested later; (2) prepare tests to obtain variations in the product similar to those which will be met with in the actual experiment; (3) adjust the difficulties of the test so that the group as a whole will discriminate between samples but some individuals will fail; (4) use test forms similar to those to be employed later; (5) start with as large a group of candidates as is feasible and with a selection test that is operationally simple if more than one stage is required; (6) screen on the basis of relative achievement, continuing until a top-ranking group of the size desired may be reliably selected; and (7) at each stage reject those who are obviously inadequate, but retain more people than will be required for the panel. This procedure is not a routine task; it requires judgment by the experimenter, particularly as to the criteria of achievement and as to how much data are needed for valid selection. According to Girardot *et al.* (1952), the multiple-stage selection assumes a good positive correlation between skills, but it will not be perfect.

It is felt that a person with previous experience might utilize some of the skill he has developed from a knowledge of techniques. Furthermore, he may note and detect differences which are unheeded by the inexperienced judge. He can often describe the sensory impressions more fully and usually has a better understanding of the particular terminology employed.

It would, however, be impossible to test independently for all of the characteristics or skills which may determine achievement. Christie (1956) believes it is not necessary. Various factors underline a unitary skill and they may be separated analytically, but in any given sensory test most of them will operate together. Realistic test situations may be set up to include acts of discrimination and judgment such as will be used later in definite experiments. Such tests will give each relevant factor its proper weight, so relative performance will be an adequate criterion for selecting the most useful panel members.

For selecting judges, Krum (1955) and Baker (1962) suggested that candidates fill out a questionnaire covering the following items: experience, availability, age, sex, health, smoking habits, quantity of particular foods habitually consumed, food prejudices, and asthmatic, physio-cardiac, and respiratory behavior. It is doubtful whether this information will be of great value; conclusive evidence against the influence of some of these factors on perception has been noted in Chapters 2 and 3. Baker's (1962) suggestion is interesting—that individuals with a physio-cardiac or asthmatic condition might be useful for certain panels since they seem to have lower thresholds for air pollutants—but the psychic attitudes of such individuals might be so unfavorable as to interfere with the tests.

Krum (1955) wrote: "It is believed that sensory ability decreases with age and that preferences change also." Therefore, he indicated, panel members should be between the ages of 20 and 50. The limiting factors are lack of experience in younger people and loss of perceptual ability in the older group. Panel members should be in good health and not physically fatigued or worried. They should not be overly susceptible to mouth and sinus infections or have frequent head colds. Persons should be eliminated who are allergic to the materials to be tested. For convenience and more accurate judging, Krum would eliminate all who do not like or refuse to eat a particular product. According to Overman and Jerome (1948), the members of the panel are frequently selected for their interest or their availability rather than for the acuity of their senses of taste and smell. In too many studies we have to "make do" with the available subjects.

C. Sensitivity Tests

In this section we discuss the many procedures that have been employed. In general, the screening tests use discrimination between solutions of known chemical composition for taste, ability to recognize odors, on-the-job performance in comparison with experienced panel members, and ability to discriminate actual differences that will be found in the

samples to be used later in the tests. The experimental situation will dictate which, if any, of these should be used.

For general panel selection we recommend that of the Quartermaster group as outlined by Girardot *et al.* (1952). In the first stage, candidates are eliminated primarily on the basis of lack of sensitivity to the sensory attributes involved, and to a lesser extent because of poor memory, slow recovery from stimulation, and failure to understand the test. In the second stage the screening is done on the basis of ability to establish and use stable subjective criteria. This double testing screens out those who will do poorly because of lack of motivation, but it does not identify in advance those who may lose interest during the course of a lengthy experiment.

Threshold tests have been used as a basis of screening by many workers. This procedure is seldom justified since there is little evidence that sensitivity to the primary tastes is related to ability to detect differences in foods. At most it is only a single factor in discriminatory ability. As King (1937) and Hopkins (1954) demonstrated, thresholds vary greatly between individuals and, except in extreme cases, no consistent relation can be demonstrated between taste acuity and palatability and judges' responses. Hall *et al.* (1959) determined the thresholds of candidates for taste and flavors on two different days, and selected those sensitive to the lowest concentrations who could duplicate their sensitivity. Hanson *et al.* (1959) used ability to detect full-strength and dilute chicken broth in selecting a panel for studying chicken flavor. A similar approach was used by Tarver *et al.* (1959), who determined for each judge a bitterness tolerance level—the recognition threshold for bitterness. Repeatability (or precision) must also be determined by standard-to-standard comparison. Hall *et al.* (1959), using that procedure, found that success in distinguishing the odd sample in triangular testing of beers showed a good correlation with the bitterness tolerance level.

Mackey and Jones (1954) tested 22 individuals to determine thresholds for primary tastes in water solutions and their ability to arrange a series in the order of concentration. Also tested was their ability to arrange, in proper order, applesauce, pumpkin, and mayonnaise containing different levels of these same taste constituents. Both the water solutions and foods could be so arranged—but the ability to arrange one properly was not highly correlated with the ability to arrange the other properly. Further, a high sensitivity did not correlate significantly with ability to arrange foods in order of concentration of taste substances. The variability among the judges was high. This experiment should be repeated.

Similar conclusions were reached by King (1937), who found no correlation between excellence in judging pure solutions and ability to rate

correctly samples of bread containing various quantities of sodium chloride, sucrose, lactic acid, and caffeine. He nevertheless suggested that the ability to identify the basic tastes at low concentration was valuable. Hopkins (1946) found a low but significant correlation between judges' ratings and the actual salt content of beef. Moreover, Krum (1955) also proposed that preliminary selection be based on sensitivity to the four primary tastes. From the results of such tests he would eliminate those who had low sensitivity. Knowles and Johnson (1941) classified judges on the basis of their sensitivity to the primary tastes but found no correlation between ability to identify the primary tastes and experience in judging foods. See also repeatability estimates of Sawyer *et al.* (1962).

Various selection tests were given to prospective panel members by Pfaffmann and Schlosberg (1952–1953), including: (1) a questionnaire designed to reveal habits, preferences, and interest in eating and drinking; (2) an odor recognition test consisting of 20 common odorous substances thought to measure interest in odors; (3) a low-odor recognition series approaching a threshold test; (4) a graded series of solutions to determine thresholds for the four primary tastes—salt, sweet, sour, and bitter; (5) use of the Elsberg blast-injection technique to determine threshold for oil of wintergreen, to detect gross departures from normal sensitivity, as from nasal obstruction; and (6) sixteen duo-trio tests on mayonnaise and thirty on an orange drink. The results failed to reveal clear evidence that any item on the questionnaire predicted performance in flavor discrimination. Selection scores on the battery of analytical tests described did not correlate well with the performance scores. The reliability coefficient (between test and retest) and the validity coefficient were very low.[*] Most noticeable was the rather unstable performance of the panel members for short-term work. No general clear-cut panel ability was evident, so that prediction of a given individual's later performance would be difficult. Those workers believe, however, that prediction of the relative ability of panel members is possible. They reported that, with the three panels tested, the score on a single discrimination session indicated who would do better on later tests: those who scored in the upper half of the total group. It is a gross measure, however, and its use might eliminate some persons who would be good performers.

[*] The words reliability and validity along with such terms as precision, accuracy, and relevance are often interpreted differently. A method of estimation which, on the average, gives the true value is called an unbiased method. Unbiased estimates are sometimes termed accurate or valid. The precision of a method refers to repeatability and is the ability of the method to produce estimates which are very close together (even if it is a biased method and is not actually measuring the true value). Thus accuracy (or validity) is related to lack of bias and precision to standard deviation.

Discrimination was measured by Morse (1954) in terms of the degree to which the individual or group can distinguish between two stimuli and communicate this distinction to investigators. Factors which affect discriminability are: (1) the individual's taste acuity at the time of the test; (2) the consistency or stability of this ability with time; (3) the distance or difference between the stimuli; (4) the design of the test, especially of its complexity and the premium it places on memory; and (5) the method of communicating the results from the subject to the investigator. Any conclusion on discriminability depends on the arbitrary standard set by the investigator of the number of correct versus incorrect judgments required. Morse required 10 correct judgments out of 12 trials for a judge to be declared discriminative, reasoning that such a ratio of judgments between equal stimuli could have occurred by chance in slightly less than 5% of similar repeated trials.

Many workers have used paired or duo-trio (Chapter 7, Section VI,A) tests for panel selection. Tarver *et al.* (1959) used a paired test for establishing bitterness tolerance levels. Byer and Gray (1953) used paired tests with beer samples, and applied χ^2 for determining the consistency of the judges. In selecting a panel for coffee testing, Harrison and Elder (1950) presented candidates with six cups of coffee consisting of three sets of pairs over a period of 20 to 30 days. The candidates were then ranked in decreasing order of their successes in making the correct pairings, and only the top half was used. Bliss (1960) used replicate paired tests with each subject. Stability of preferences was used as the selection criterion. Lockhart (1951) noted that any of the binomial systems provides a means for rapidly selecting panel members whose sensitivities can be described in terms of probability levels. These systems can also be used for checking the sensitivities of the panel on a day-to-day or week-to-week basis.

The most common method of choice has been the triangle test (Chapter 7, Section VI,B). It was first used by Bengtsson and Helm (1946) and Helm and Trolle (1946) for selecting beer tasting panels. Beers of known differences were used first in simple tests and later in more difficult tests. Only the most sensitive individuals were used. Data from the tests were used to check panel performance. The Quartermaster group (Girardot *et al.*, 1952) used a triangle test in the first stage of selection. Simple tests were used first, but later the tests were of increasing difficulty. The judges were ranked on the basis of their percentages of correct judgments. All judges took about the same number of tests at each level of difficulty. Only the ranking near the cut-off point is critical.

Bradley (1955) recommended repeated triangle tests for selecting judges. Sequential methods (Chapter IV, Section III) can be recom-

mended because of their efficiency and because they focus attention on the risk of accepting poor judges or of rejecting good ones. Using both paired and triangle tests, Schlosberg *et al.* (1954) found that a judge's relative performance during the first two days of testing "had a fair predictive value for his relative over-all performance during the following 20-day period." This was not true when preference for milk was measured, but that result will be discussed later (Chapter 6, Section II,F). Their experience was that ability for one panel did not carry over to another. Hening (1948) used the triangle test to select panels for distinguishing differences in flavor resulting from time and temperature of storage of various products. Amerine (1948) recommended it for selecting wine panels. Krum (1955) likewise used it, noting that each candidate should take the same number of tests. The cut-off point was determined by the number of panel members required and the precision required by the problem. Moser *et al.* (1950) found one experienced judge with an excellent record in testing oil but a poor record in detecting diacetyl by triangle tests. They attributed this disparity to confusion on the part of the subject. However, this judge may have been insensitive to low concentrations of diacetyl, even though reputed to have a keen sense of smell.

Dawson *et al.* (1963) showed that for taste thresholds the paired comparison resulted in lower thresholds than the triangular, and that the single-sample procedure was the least sensitive.

Various methods of scoring have been used in selecting panels. Hedonic scores were used by Girardot *et al.* (1952). Similar procedures have been used or reviewed by Sharp *et al.* (1936), Trout and Sharp (1937), Boggs and Hanson (1949), Harrison *et al.* (1954), and others. Used to evaluate performance have been average deviations between duplicate scores, the deviation from the score of a control sample introduced in series, or the deviations of scores between first and second tastings (with the samples coded and presented in different orders). Although these measure individual reproducibility, they do not relate reproducibility with one sample to ability to find differences between unlike samples. To rectify this, the correlation coefficient between the first score and duplicate scores for a series of samples of varying quality may be used. Bennett *et al.* (1956) used the standard error of the means to measure ability to reproduce judgments. Hopkins (1946) calculated both correlation coefficients and regression equations to relate each judge's assessment to the average of the panel. A range of sensitivity was demonstrable and the suitability of individuals for tests could be evaluated. The correlation coefficients were much higher for biscuits than for dried milk. Moser *et al.* (1950) likewise calculated the correlation co-

efficient and regression equation. Used for selection was the standard error of regression of the individual's scores with the average of the whole panel.

Overman and Jerome (1948) used scores and applied two tests: a comparison of the average range of a judge's scores and a comparison of the number of times a judge duplicated his score. The first, being rapid and easy to understand, was preferred for preliminary evaluation. Deviation from the mean was the statistical measure employed. A high deviation from an individual's mean indicates inability to duplicate judgments or marked changeability of opinion during the tests. A low deviation from his own mean indicates either a high degree of reproducibility or a lack of critical discrimination (as when all scores are very high or very low). Since the method of deviation from the mean may obscure discrimination, Overman and Jerome preferred analysis of variance for determining the consistency of the judges and their ability to discriminate. The variance ratios (F values) were used as an index of a judge's ability to duplicate his scores and can be used as an appropriate measure of the consistency of each judge. If members of the panel show marked differences in individual-error variances it is advisable to test the panel for homogeneity of variances before comparing individuals. Krum (1955), for his panel, also selected judges with highly significant F ratios (Chapter 10, Section V,A). Girardot *et al.* (1952) employed analysis of variance and demonstrated the superiority of the selected panels. Wiley *et al.* (1957) screened judges on the basis of F ratios. Those with a statistical significance at odds of 9:1 were retained. They were then retested, and only those were selected whose F values indicated significance at odds of 19:1. The quick method of "range ratio" of Tukey (1951) was also employed. In this method, if the ratio of (range among treatment) to (range within treatments $\times$ factor) $= 1$, then the difference is significant. (The factor depends on the number of treatments and is obtained from a table.)

Sawyer (1958) and Sawyer *et al.* (1962) based panel selection on repeatability—the interclass correlation of repeated measurements (a measure of the constancy of repeated observations by a given judge). This is a point estimate, and is estimated directly from variance analysis of discrimination test data. The proportion of judges whose sensitivity satisfies established specifications can be predicted. In these studies the average repeatability of performance was equivalent to or greater than the repeatability predicted by analysis of variance. "Thus, estimates of intraclass correlation appear to provide a reliable basis for predictions in the selection of panels" (Sawyer *et al.*, 1962).

Simple ranking of judges' scores often permits relative differentiation

of individual capabilities but does not ensure a specified level of proficiency.

Kramer (1955) recommended choosing judges on the basis of their ability to detect differences at a given probability level. His procedure involved matching concentrations, and the tables he published should be useful whether or not duplicates are available for all samples.

Probably because of their extensive use in industry, control charts have been used in selecting panels or maintaining level of performance. A control chart is a statistical device used principally for the study and control of repetitive processes. Such charts are based on the theory that variations due to chance occur in a random pattern and that the frequencies approach those of the binomial distribution. To see whether a process is out of control, past data are plotted on a control chart. If the data conform to a pattern of random variation within the control limits, the process will be judged as being in control. Reliability is indicated by the narrowness of spread between control limits. Since pre-established standards can be set up, the control chart also measures the validity of the judge's results. For basic data, see Feigenbaum (1951) and Duncan (1959).

Control charts have been recommended by Marcuse (1945, 1947), Moser *et al.* (1950), Harrison and Elder (1950), Krum (1955), Coote (1956), and Tarver and Ellis (1961). With them, not only an individual's performance but that of an entire panel can be held to a given precision.

Harrison *et al.* (1954) defined the efficiency of a panel in terms of the probability of the panel's acceptance of definite differences in the samples. To eliminate the number of correct selections through chance alone, the scores were corrected with the following formula:

$$S = \frac{100(R - C)}{100 - C}$$

where S is the percent score corrected for chance expectation, R the raw percent score, and C the percent score expected by chance.

More elaborate mathematical procedures may be used in certain cases: multiple-factor analysis, item analysis, discriminate functions, product-moment correlation coefficients (Filipello, 1957), etc.

In most cases a simple test using some binomial procedure may be used to eliminate insensitive judges. See Amerine *et al.* (1959) for detailed procedures used for wine panels. Analysis of variance or some sequential procedure should be used for more complex situations or to maintain the panel at some desired level of performance.

Variation among 30 judges in scoring scrambled eggs containing various amounts of added primary-taste compounds was described by Hop-

kins (1946). Significant
Some statistically sign
containing different te
substances was also fo
erratic as quality det
relation between taste
anticipated. Quality c
sensations as well as t
scoring methods used
judges (see also Chap

Sensitivity to taste
discrimination. In mos
necessary since absolu
lated to perceptual skil

D. PANEL SIZE

The number of jud
cording to the variabili
liminary experiment wi
the number of judges
significance. As quality
panel size must be i
tically significant (Bog
A good example of thi
biscuits, dried eggs, bu
levels of acceptability
judges were required
is available, however,
crimination. In incom
surprisingly, that the c
intermediate quality.

Of course, the pane
in difference testing. H
ing in degree of saltine
nate sensory difference
panel comparisons, 62 j
preferred panels of 30
Helm (1946) preferred
which might influence
were believed adequate
When only three or fo
sible to repeat the tests

kins (1946). Significant variation ($p = 0.01$) was observed among judges. Some statistically significant discrimination among groups of samples containing different test substances and among concentrations of these substances was also found. Individual scores became progressively more erratic as quality deteriorated. Hopkins concluded that no consistent relation between taste acuity alone and palatability judgments should be anticipated. Quality evaluation includes visual, olfactory, and tactile sensations as well as taste sensitivity, and is further conditioned by the scoring methods used and the experience and frame of reference of the judges (see also Chapter 8).

Sensitivity to taste or odor appears to be only one factor influencing discrimination. In most cases, elaborate tests based on acuity seem unnecessary since absolute sensitivity to the basic tastes is not closely related to perceptual skills.

D. Panel Size

The number of judges needed in a given experiment will vary according to the variabilities of the individuals and of the product. A preliminary experiment will give information from which can be calculated the number of judges necessary to secure a given level of statistical significance. As quality decreases, variability among judges increases and panel size must be increased to obtain differences which are statistically significant (Boggs and Hanson, 1949; Kefford and Christie, 1960). A good example of this is found in work by Hopkins (1946, 1947) with biscuits, dried eggs, butter, dried milk, and bacon. He noted that, at low levels of acceptability, discrimination was very erratic, so that more judges were required for significance in results. Not enough information is available, however, on the interrelationship of acceptability and discrimination. In incomplete-block studies, Hanson *et al.* (1951) found, surprisingly, that the error of the panel means was greater for samples of intermediate quality.

Of course, the panels must be much larger in preference testing than in difference testing. Hopkins (1947) concluded that, with bacon varying in degree of saltiness, panels of 35 judges were necessary to discriminate sensory differences of 5% with intrapanel comparisons. For interpanel comparisons, 62 judges would be necessary. Girardot *et al.* (1952) preferred panels of 30 to 90 in food-development studies. Bengtsson and Helm (1946) preferred large panels (50 to 100) in testing for differences which might influence future work. For routine control, 10 to 30 judges were believed adequate. Krum (1955) found panels of 10 to 30 sufficient. When only three or four individuals were available he believed it possible to repeat the tests enough times to get a suitable number of results.

SENSORY ANALYSIS OF FOODS

Edited by

J. R. PIGGOTT

Department of Bioscience and Biotechnology, Food Science Division, University of Strathclyde, Glasgow, Scotland, UK

ELSEVIER APPLIED SCIENCE PUBLISHERS
LONDON and NEW YORK

ELSEVIER APPLIED SCIENCE PUBLISHERS LTD
Ripple Road, Barking, Essex, England

Sole Distributor in the USA and Canada
ELSEVIER SCIENCE PUBLISHING CO., INC.
52 Vanderbilt Avenue, New York, NY 10017, USA

British Library Cataloguing in Publication Data

Sensory analysis of foods.
1. Food—Sensory evaluation
I. Piggott, J. R.
664'.07 TX546

ISBN 0-85334-272-5

WITH 68 TABLES AND 60 ILLUSTRATIONS

The selection and presentation of material and the opinions expressed in this publication are the sole responsibility of the authors concerned.

Printed in Great Britain by Galliard (Printers) Ltd, Great Yarmouth

PREFAC

In 1965, a book which has since occupie
science of sensory analysis was publishe
R. M. and Roessler, E. B., *Principles*
(Academic Press, New York). The author
also hope that it will stimulate further rese
can be seen from the rapid growth in the l
evaluation of foods and beverages. Since
senses has grown; new sensory methods
been improved, both in application
powerful computers are widely available

Reviews of these developments are not
in compiling this book was to provide a
knowledge and practice in sensory analys
a laboratory manual, but to provide the p
review of progress throughout the work
foundation on which to build his own ex

Individual chapters have been contribu
their fields, who have surveyed and interp
chapters are concerned with examinations
provide a basic understanding of the se
sensory testing and by which food flavou
These are followed by descriptions of speci
appearance assessment, and by reviews c
ranking and scaling methods, and descrip
laboratory, a chapter is devoted to con
concluded by descriptions of the sta
descriptive and inferential analysis of sens

TABLE 4
RULES FOR SCALE DESIGN

a. An entire scale should use one root word, e.g. like—like/dislike. Do not shift from prefer to dislike.
b. Every scale point should be modified with modifiers of the root such as very, slightly, etc.
c. There should be the same number of scale values above and below neutral, if neutral is used, or above and below the midpoint. Do not use five levels of like and three of dislike.
d. Carefully consider the use of the neutral points (e.g. neither like nor dislike); use one if it is logically necessary.
e. Use an adequate number of scale points; err in the direction of too many scale points.

develop his/her own scales only rarely! It is preferable and safer to use scales which you have used previously with demonstrated success (defined statistically) or which have been used and demonstrated by others. The most well-known scale in food research is the nine-point hedonic scale (Table 5) developed by the US Army in the 1940s. It is interesting to note that this scale satisfies the five points mentioned above: it is adequately long (nine points), it possesses a neutral point, it uses one root word (like–dislike), and uses the same modifiers above and below neutral (slightly, moderately, very and extremely).

Bass *et al.* (1974) have scaled verbal descriptors of frequency (Table 6) and amount (Table 7), both key concepts in food attitude research. These data provide guidelines for selecting four–nine point category scales for these concepts. The four-point scale of frequency (never, sometimes, often

TABLE 5
THE NINE-POINT HEDONIC SCALE USED FOR FOOD ACCEPTANCE AND FOOD PREFERENCE

9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

TABLE 6
SCALES OF FREQUENCY
(Adapted from Bass *et al.*, 1974)

9	8	7	6	5	4
Always	Always	Always	Always	Always	Always
[illegible]	Continually	Constantly	Frequently if not always	Very often	Often
				[illegible] times	Sometimes

with questions of obesity and diet, with questions of how much money is spent on food, and whether a parent prepares wholesome meals for the children. Special care is needed when such things are asked, as well as anytime the respondent perceives that he might lose or gain something as a result of the survey.

10. Periodical behaviour. Asking people how often they eat out of doors, how often they prefer to drink cold beverages, etc., might involve behaviour which varies with the season.

11. Memory. When the surveyor is depending on the memory of the respondent, it is best to assume that memory will not be as good as we would like to think it is. In many cases it has been demonstrated that memory is very poor for seemingly simple information. Beware.

2.3. Food Preference

The nine-point hedonic scale of food preference has been the one most commonly used. It was developed by the Quartermaster Food and Container Institute (QMFCI) in Chicago in the late 1940s (see Peryam and Pilgrim, 1957).

In addition, the respondent is permitted to indicate that a food was 'never tried'. The development of this scale involved more research than other food preference measures. A rating scale was selected rather than a paired-comparison method, in which pairs of items are used rather than lists of individual items. It was determined that the rating scale approach and the paired-comparison approach yielded relative preferences in good agreement.

The two basic questions addressed were the number of scale points and their labelling. The nine-point scale had already been used in laboratory food acceptance testing, and the researchers compared it in a food preference testing (using 50, 100 and 150 item lists) to a seven-point scale (eliminating the 'like extremely' and 'dislike extremely' categories) and a five-point scale (eliminating the 'extremely' and 'slightly' categories). The three survey lengths showed no difference in test–retest reliability and the nine-point category showed the highest test–retest reliability (0·96), so the longer list length and the longer scale length were both adopted.

The naming of the scale points was the next step in the development of the scale. Ideally, the names should be chosen by scaling their meaning, so that the distance between 'extremely' and 'very much' should be the same as that between 'very much' and 'moderately'.

One subtle point involved in using rating scales is positioning on the page. In their survey the rating scale followed a list of foods presented on

the left-hand side of the paper. The question r
begin with the 'dislike' or 'like' end of the rating s
on a list of 45 food items, that the proportion of
categories was almost identical (correlation c
there were significant differences between the
'dislike extremely' was placed on the extreme lef
extremely' was placed in that position. Beginr
extremely' led to a significantly greater frequenc
Beginning the scale with 'like extremely' did n
increased frequency for 'like' categories. In pra
very small. The correlation between the 45 pair
and it is the mean which is used for predictive pu
researchers suggested that the scale should beg
hastened to add that no clear problem resulte

The issue of preference frequency has be
preference scaling and has been phrased in a
would you like to eat the menu items?', 'How o
eat the items?', 'How often would you like t
'How often would you like to eat the item?'.

Preference frequency scales have been of tw
categories of frequency and the other using qua
8). Almost all frequency scales used have had f
verbal-based scales have depended heavily on th
of day, week and month. Two scales have used
to day, week or month referrents (Leverton, 194
this could represent difficulties in trying to tran
units. Benson (1958) also used a four-category s
terms (once a day, week, month, year). Hartmu
et al. (1967) used identical nine-category scales f
month' (plus 'never want'). The QMFCI researc
used a nine-category scale which overlapped gre
In some administrations it was extended to 'e
'once a year'. The question which arises th
appropriate time frame for the preference freq
has not been directly addressed, most scales u
For most purposes it would appear that items e
would be insignificant, unless very specialis
restaurants, catering) were of interest.

The QMFCI scale also listed the frequency p
categories. The category 'every three month

questions of how much money is
repares wholesome meals for the
such things are asked, as well as
: might lose or gain something as a

:ople how often they eat out of
old beverages, etc., might involve

depending on the memory of the
emory will not be as good as we
:s it has been demonstrated that
ple information. Beware.

›reference has been the one most
y the Quartermaster Food and
› in the late 1940s (see Peryam and

d to indicate that a food was 'never
:volved more research than other
was selected rather than a paired-
tems are used rather than lists of
the rating scale approach and the
relative preferences in good

re the number of scale points and
l already been used in laboratory
:archers compared it in a food
0 item lists) to a seven-point scale
slike extremely' categories) and a
ely' and 'slightly' categories). The
:e in test–retest reliability and the
test–retest reliability (0·96), so the
:ngth were both adopted.
e next step in the development of
iosen by scaling their meaning, so
'very much' should be the same as
:ly'.
ating scales is positioning on the
owed a list of foods presented on

the left-hand side of the paper. The question raised was 'Should the scale begin with the 'dislike' or 'like' end of the rating scale?'. They found, in a test on a list of 45 food items, that the proportion of answers in each of the nine categories was almost identical (correlation coefficient = 0·96). However there were significant differences between the form of the scale in which 'dislike extremely' was placed on the extreme left and the one in which 'like extremely' was placed in that position. Beginning the scale with 'dislike extremely' led to a significantly greater frequency of the 'dislike' categories. Beginning the scale with 'like extremely' did not produce the analogous increased frequency for 'like' categories. In practical terms the effects are very small. The correlation between the 45 pairs of food means was 0·997, and it is the mean which is used for predictive purposes with these data. The researchers suggested that the scale should begin with 'like extremely' but hastened to add that no clear problem resulted from the reverse.

The issue of preference frequency has been another focus of food preference scaling and has been phrased in a variety of ways: 'How often would you like to eat the menu items?', 'How often would you be willing to eat the items?', 'How often would you like to see the food offered?'..., 'How often would you like to eat the item?'.

Preference frequency scales have been of two types, one using verbal categories of frequency and the other using quantitative categories (Table 8). Almost all frequency scales used have had four or nine categories. The verbal-based scales have depended heavily on the existing temporal system of day, week and month. Two scales have used the term 'often' in addition to day, week or month referrents (Leverton, 1944; Schuck, 1961), although this could represent difficulties in trying to translate into actual temporal units. Benson (1958) also used a four-category scale but stuck to temporal terms (once a day, week, month, year). Hartmuller (1971) and Knickrehm *et al.* (1967) used identical nine-category scales from 'twice a day' to 'once a month' (plus 'never want'). The QMFCI research on frequency scales also used a nine-category scale which overlapped greatly with the one just cited. In some administrations it was extended to 'every three months' and to 'once a year'. The question which arises then is: 'What is the most appropriate time frame for the preference frequency scale?' This question has not been directly addressed, most scales using the month as the unit. For most purposes it would appear that items consumed only once per year would be insignificant, unless very specialised food services (class A restaurants, catering) were of interest.

The QMFCI scale also listed the frequency per month of all verbal scale categories. The category 'every three months' was rated 0·3 and the

TABLE 8
SCALES OF PREFERRED FREQUENCY

	Knickrehm et al. (1967)	*Hartmuller (1971)*	*QMFCI*	*Benson (1958)*	*Schuck (1961)*	*Leverton (1944)*
Often					*	*
Twice a day	*	*	*			
Once a day	*	*	*	*		
Every other day	*	*				
Several times per week, 15 months			*			
Twice a week	*	*	*			
Once a week	*	*	*	*	*	*
Every other week	*	*	*			
Once a month	*	*	*	*		
Every 3 months			*			
Once a year			*	*		
Never/unwilling to eat	*	*	*			*
Not familiar	*	*	*		*	*

category 'once a year' was rated 0·1. This reinforces the use of the month as the unit. It also provides both the test respondent and the researcher with a quantified scale for analysis and prediction. In some cases (e.g. Knickrehm *et al.*, 1967) subjects responded on the frequency scale by listing the number of the verbal category. For example, twice a week was coded as 4. The potential problem here is that the best respondent is not using the actual frequency statement in his answer, whereas in other scales he is.

A preference scale (Fig. 1) developed more recently for the military used a quantitative preference frequency scale based on the week and month (Meiselman *et al.*, 1972). The subject was asked how often he would like an item in terms of days per week (answer 1, 2, 3, 4, 5, 6 or 7) and weeks per month (answer 1, 2, 3 or 4). While this does directly ask the preference frequency question in quantitative terms, it forces the subject into a week month system. If he wants squash 13 times per month, he cannot so indicate. Further it assumes that the weekly pattern is repeated. This is also the case in some verbal categories scales. A more recently developed survey (Fig. 2) (Meiselman and Waterman, 1978) avoids weekly units and asks for preference frequency per month using a scale which permits coding of any number from 0 to 31 (actually 39 is possible) days per month. Note again that the monthly unit was the unit of choice.

The numerical and verbal scales possibly
the subject is using numbers in the numeri
categories in the non-numerical scale. Whe
codes for the verbal scale categories, proble
attention is then on a number which does no
He then begins to use the category scale of
referring them to their referrent frequencie
happen in the hedonic acceptance scale in wh
without realising its referrent (extremely go

One potential advantage of certain quan
that they can be ratio scales, that is, scales v
point. Ratio scales permit statements of ra
preferred twice as often as *y*, etc. The freq
Army Natick Laboratories (Meiselman ar
scale (from 0 to 39). Both the old QMFC
Meiselman *et al.* (1972) are not continuou
subject is selecting categories rather than

The scales discussed so far have been cit
frequency scales. Schutz (1965) developed a
Scale), by scaling 18 action statements
towards foods. Nine were selected to give
deviation and mean of the FACT scale and
very similar; the two scales correlate 0·9
tendency for the FACT means to be lo
apparently results from slightly lower F
semisolid and liquid foods.

Van Riter (1956) used a scale based on
vegetables) including scale categories: 'neve
of my family dislike the food', and 'prepa
categories are indicators of factors that
preference determination. Whether the
preferences themselves is unclear without

2.4. Examples of Food Preference Data

Although a large amount of food prefer
institutions and commercial organisatio
literature. However, there is a growing bo
tap so that many food preference decisi
One of the largest of available data bases
Forces which have been collecting food pr

The numerical and verbal scales possibly reduce to the same thing when the subject is using numbers in the numerical scale and using the verbal categories in the non-numerical scale. When the subject uses numerical codes for the verbal scale categories, problems can arise. The focus of his attention is then on a number which does not directly represent frequency. He then begins to use the category scale of numbers without necessarily referring them to their referrent frequencies. This is similar to what can happen in the hedonic acceptance scale in which one begins to use a number without realising its referrent (extremely good, very bad, etc.).

One potential advantage of certain quantitative scales of frequency is that they can be ratio scales, that is, scales with equal intervals and a zero point. Ratio scales permit statements of ratios so that one could say x is preferred twice as often as y, etc. The frequency scale developed by US Army Natick Laboratories (Meiselman and Waterman, 1978) is such a scale (from 0 to 39). Both the old QMFCI scale and the scale used by Meiselman *et al.* (1972) are not continuous series of numbers; hence the subject is selecting categories rather than dealing in ratios.

The scales discussed so far have been either hedonic scales or preference frequency scales. Schutz (1965) developed a food action rating scale (FACT Scale), by scaling 18 action statements representing affective attitudes towards foods. Nine were selected to give equal intervals. The standard deviation and mean of the FACT scale and the nine-point hedonic scale are very similar; the two scales correlate 0·97 for food means. The overall tendency for the FACT means to be lower than the hedonic means apparently results from slightly lower FACT ratings for desserts and semisolid and liquid foods.

Van Riter (1956) used a scale based on home use of foods (specifically vegetables) including scale categories: 'never served at home', 'one or more of my family dislike the food', and 'prepared differently at home'. These categories are indicators of factors that are possibly important in food preference determination. Whether they are good measures of the preferences themselves is unclear without a more complete evaluation.

2.4. Examples of Food Preference Data

Although a large amount of food preference data is collected by various institutions and commercial organisations, little of it reaches the open literature. However, there is a growing body of data for the investigator to tap so that many food preference decisions need not be made intuitively. One of the largest of available data bases is that of the United States Armed Forces which have been collecting food preference data for almost 40 years.

REQUENCY

MFCI	Benson (1958)	Schuck (1961)	Leverton (1944)
		*	*
*			
*	*		
*			
*			
*	*	*	*
*			
*	*		
*			
*	*		
*			*
*		*	*

nforces the use of the month as ndent and the researcher with a In some cases (e.g. Knickrehm :ncy scale by listing the number : a week was coded as 4. The ›ondent is not using the actual s in other scales he is.

'e recently for the military used ›ased on the week and month ked how often he would like an , 3, 4, 5, 6 or 7) and weeks per ›es directly ask the preference it forces the subject into a times per month, he cannot so pattern is repeated. This is also nore recently developed survey voids weekly units and asks for e which permits coding of any :) days per month. Note again :e.

role in the fitting algorithm. Following Carroll (1972), it is necessary to distinguish two modes of analysis. In *internal* analysis the objective is to achieve a consensus configuration of the stimuli based solely on the preference data. In external analysis the aim is to relate preferences to physicochemical measurements using as parsimonious a model as possible to take account of individual differences in scoring patterns.

3.3.2. Internal Analyses

The simplest approach to modelling individual differences in preference is the vector model proposed by Tucker (1960). The set of stimulus points are embedded in a multidimensional space and each subject is represented by a vector in the space. The ordering of the projections of the stimulus points on to the vector gives the preference ranking of that subject. The cosine of the angle that a vector makes with the dimensions of the space is considered to be proportional to the relative importance of that dimension in the preference judgement.

An example from our own experience demonstrates the use of the vector model very effectively. The data (unpublished) was generated at the Torry Research Station, Aberdeen, and we are grateful to P. Howgate for permission to use it. In this study, 48 subjects were asked to rate six types of fish or fish product on an hedonic (Peryam/Pilgrim) scale: 1 = dislike extremely, 9 = like extremely. For brevity Table 4 shows the session means for only six subjects, A–F. The complete data was input to the MDPREF program (Chang and Carroll, 1968), and the two-dimensional solution, which accounts for 85·3 % of the variation, appears in Fig. 9. The subjects appear as points on the unit circle and a preference ranking is obtained by

TABLE 4

MEAN PREFERENCE SCORE FOR SIX SUBJECTS ON SIX FISH PRODUCTS

Subject	*Stimulus*					
	White fish in parsley sauce	*White fish fingers*	*Scad (good)*	*Scad (poor)*	*Cod mince fingers*	*Blue whiting fingers*
A	7·7	7·3	7·4	6·0	7·3	8·0
B	5·0	5·2	6·2	3·7	5·3	5·0
C	7·5	6·6	5·0	4·0	6·3	4·5
D	7·8	6·8	5·6	4·7	6·0	5·7
E	6·5	6·3	6·1	6·5	5·3	3·9
F	5·7	6·8	6·0	6·6	7·2	4·7

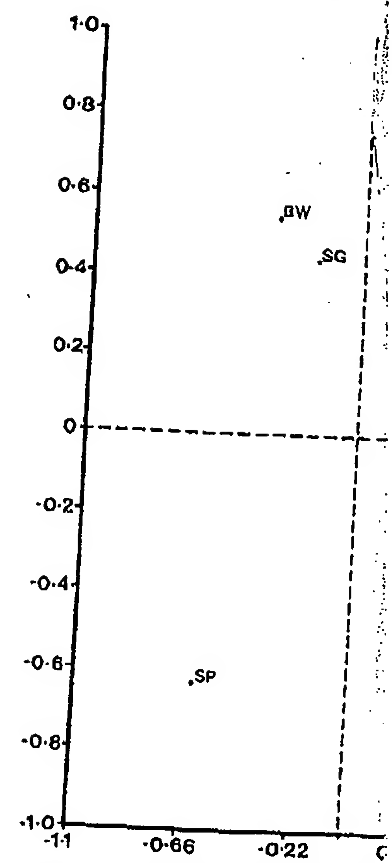


FIG. 9. MDPREF solution displaying configu fish in parsley sauce; WF, white fish fingers; SG, mince fingers; BW, blue whiting. Subjects are di the unit circle

drawing a line passing through the orig perpendiculars from each stimulus point o

The horizontal dimension, around which the conventional preference dimension. The this dimension in the expected order and the the left with the reformed product of p acceptability score. However, there are sub

STATISTICAL METHODS IN FOOD AND CONSUMER RESEARCH

MAXIMO C. GACULA, JR.
Department of Biostatistics
Armour Research Center
Armour–Dial, Inc.
Scottsdale, Arizona

JAGBIR SINGH
Department of Statistics
School of Business Administration
Temple University
Philadelphia, Pennsylvania

1984

ACADEMIC PRESS, INC.
(Harcourt Brace Jovanovich, Publishers)
Orlando San Diego New York London
Toronto Montreal Sydney Tokyo

To Mely, Karen, Lisa, Elen
Veena, Rajesh, Seema, Sap

ACADEMIC PRESS, INC.
Orlando, Florida 32887

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

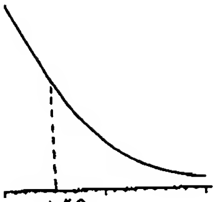
Library of Congress Cataloging in Publication Data

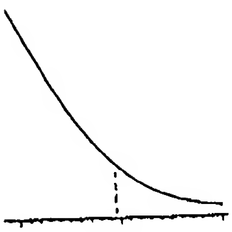
Gacula, Maximo C.
Statistical methods in food and consumer research.

(Food science and technology)
Includes index.
1. Food research--Statistical methods. 2. Consumers--
Research--Statistical methods. 3. Experimental design.
I. Singh, Jagbir. II. Title.
TX367.G33 1985 664'.0072 84-11170
ISBN 0-12-272050-4 (alk. paper)

PRINTED IN THE UNITED STATES OF AMERICA

84 85 86 87 9 8 7 6 5 4 3 2 1





urve. Ordinate, $f(X)$.

$1 - P(Z \leq -1.960)$
5

$- P(Z < 0.500)$

The pdf of the standard normal distribution, denoted by $\phi(X)$, is readily seen from (1.1-5) when $\mu = 0$ and $\sigma^2 = 1$. That is,

$$\phi(X) = (1/\sqrt{2\pi})\exp(-\tfrac{1}{2}X^2), \tag{1.1-8}$$

where $-\infty < X < \infty$. If Z is $N(0, 1)$, the cumulative probability $P(Z \leq X)$ is denoted by $\Phi(X)$ and is called the cumulative distribution function (cdf) of the standard normal r.v. Z.

In statistical inference we also come across three other probability distributions. They are called the t distribution, the chi-square (χ^2) distribution, and the F distribution. Both the t and χ^2 distributions depend on only one parameter, whereas the F distribution depends on two. In statistical terminology, the parameters of these distributions are called the degrees of freedom (DF) parameters. For the F distribution the two parameters are identified as the "numerator" and the "denominator" DF. The percentiles of the t distribution are given in Table A-2. As the DF increase, the percentiles of the t distribution approach those of the standard normal distribution. That is, for large DF (>30) the t distribution can be approximated by the standard normal distribution. Some selected percentiles of the χ^2 distribution for various values of the DF parameters are given in Table A-3. Frequently used percentiles of the F distribution for various combinations of the numerator and the denominator DF are given in Table A-4.

Throughout this book we shall use Z_α, $t_{\alpha,v}$, $\chi^2_{\alpha,v}$, and F_{α,v_1,v_2}, to denote, respectively, the αth percentile of the standard normal distribution, the t distribution with v DF, the χ^2 distribution with v DF, and the F distribution with the numerator DF v_1 and the denominator DF v_2. It is useful to know that

$$1/F_{\alpha,v_1,v_2} = F_{1-\alpha,v_2,v_1}.$$

There are other probability distributions, such as the lognormal, the Weibull, the binomial, and the exponential, which will be introduced whenever needed. The probability background discussed here should suffice as a beginning.

Estimation

As mentioned, the population parameters are usually unknown and are to be estimated by appropriate sample quantities called *statistics*. For example, statistic $\bar{X}$, the sample mean, may be used to estimate population mean μ. Also S^2, the sample variance, can be used to estimate population variance σ^2. A statistic, when it is used to estimate a parameter, is called an *estimator*. Since an estimator is a sample quantity, it is subject to sampling variation and sampling errors. That is, the values assumed by an estimator vary from one sample to another. In this sense, the possible range of values of

an estimator is governed by a chance or probability model, which quantifies the extent of sampling variation and sampling error. If an estimator $\hat{\theta}$ for estimating θ is such that the mean of the distribution of $\hat{\theta}$ is θ, then $\hat{\theta}$ is called an *unbiased* estimator. In statistical language the mean of the distribution of $\hat{\theta}$ is also called the *expected value* of $\hat{\theta}$ and is denoted by $E(\hat{\theta})$. In this notation, an estimator $\hat{\theta}$ of θ is unbiased if $E(\hat{\theta}) = \theta$. Both the sample mean $\bar{X}$ and variance S^2 are unbiased estimators of the population mean μ and variance σ^2, respectively. That is, $E(\bar{X}) = \mu$, and $E(S^2) = \sigma^2$. However, sample standard deviation S is not an unbiased estimator of the population standard deviation σ. In our notation, $E(S) \neq \sigma$. But there are other criteria in addition to the unbiasedness, such as consistency or efficiency, which may be used in deciding appropriate estimators. We need not go into the theory and details because that is not the aim of this book, but we shall use only the statistically established "best" estimators of the parameters of concern to us.

An estimator when used to estimate a parameter by just one number is called a *point estimator*, and the resulting estimate is called a *point estimate*. Similarly, if a parameter is estimated by a range of values, then we have an *interval estimate*.

For estimating population mean μ by an interval estimate, we can specify an interval, for example, $(\bar{X} - S, \bar{X} + S)$, for the likely values of μ, where S is the sample standard deviation. Obviously, the range of an interval estimate may or may not contain the true parameter value. But we can ask: How confident are we in using interval $(\bar{X} - S, \bar{X} + S)$ to contain the true value of μ, whatever that is? To answer this sort of question, statisticians use the so-called *confidence intervals* for estimation. For example, the degree of confidence is about $(1 - \alpha)100\%$ that the following interval,

$$\bar{X} - t_{\alpha/2,n-1}(S_{\bar{X}}), \quad \bar{X} + t_{\alpha/2,n-1}(S_{\bar{X}}),$$

contains population mean μ, where $S_{\bar{X}} = S/\sqrt{n}$ is an estimate of the standard deviation or the standard error of $\bar{X}$. We may emphasize here the $\bar{X}$ is a sample quantity and, therefore, is subject to sampling variations and sampling errors. It is the quantification of the sampling variation and errors in the distribution of $\bar{X}$ that lets statisticians declare the degree of confidence associated with an interval estimate based on $\bar{X}$. A very useful result pertaining to the probability distribution of $\bar{X}$ is the following.

Consider a population with mean μ and variance σ^2. Suppose that we can list all possible random samples of size n from this population and compute $\bar{X}$ for each sample, thus generating the distribution of $\bar{X}$. Theoretical statisticians have shown (the central limit theorem) that the distribution of $\bar{X}$, for all practical purposes, is normal with mean μ and variance σ^2/n. That is, $\bar{X}$ is $N(\mu, \sigma/\sqrt{n})$.

Testing of Hypotheses

Another area of statistical infer theses. The testing of hypotheses con

1. Formulation of hypotheses
2. Collection, analysis of data,
3. Specification of a decision ru or rejecting hypotheses

The formulation of the hypo proposed experiment. For instance whether a process modification h researcher proceeds by producing t modified process. Let μ denote the while μ_0, the mean texture of th *hypothesis* is written as

H_0:

which states that on the average t modified process. The *alternative h*

H_a

which states that there is a change process. The alternative H_a in (1.1 may be less than μ_0 $(\mu < \mu_0)$ *alternative* hypothesis is either

H_a: $\mu < \mu_0$

The formulation of a one-sided h of the experimenter.

If, instead of in the mean, similarly formulate null and alter alternative hypotheses have been used to develop statistical decisic

Since a statistical decision account for sampling variations of the null hypothesis on the ba rejection of the null hypothesis the probability of which is de $\alpha = 0.05$ indicates that on the a hypothesis 5 times in 100 cases. the statistical test; values of 0.